CLAIMS

What is claimed is:

1. A method for performing a record append operation in a system that includes a primary replica and one or more secondary replicas, comprising:

receiving, at the primary replica, a record append request that identifies a record to be appended to a chunk, the one or more secondary replicas storing copies of the chunk;

determining whether the record fits into the chunk;

appending the record to the chunk when the record fits into the chunk;

forwarding the record append request to the one or more secondary replicas; and

appending the record to the copies of the chunk by the one or more secondary replicas.

2. The method of claim 1, wherein the determining whether the record fits into the chunk includes:

determining whether appending the record to the chunk would cause the chunk to exceed a predetermined size.

3. The method of claim 1, further comprising:

padding the chunk to a predetermined size when the record does not fit into the chunk.

4. The method of claim 3, further comprising:

notifying the one or more secondary replicas to pad the copies of the chunk to the predetermined size.

5. The method of claim 1, further comprising:

notifying a sender of the record append request to retry the record append request on a next chunk when the record does not fit into the chunk.

6. The method of claim 1, wherein the forwarding the record append request includes:

identifying an offset at which the record was appended to the chunk.

7. The method of claim 6, wherein the appending the record to the copies of the chunk includes:

writing the record at the identified offset in the copies of the chunk.

8. The method of claim 7, further comprising:

determining that the record append operation is successful when the record is written at the identified offset in the chunk and the copies of the chunk.

9. The method of claim 1, wherein at least one of the appending the record to the chunk and the appending the record to the copies of the chunk includes:

writing the record as an atomic unit.

10. The method of claim 1, wherein the receiving a record append request includes:

concurrently receiving a plurality of record append requests, and

serializing the plurality of record append requests.

11. The method of claim 1, wherein the primary replica concurrently processes a plurality of record append requests.

12. The method of claim 1, wherein the system further includes a master and a plurality of servers, the method further comprising:

receiving a request for identification of the primary replica,

determining whether any of the servers have a lease on the chunk, and

identifying one of the servers as the primary replica when the one server has a lease on the chunk.

13. The method of claim 1, further comprising:

receiving, from a sender of the record append request, the record to be appended.

14. The method of claim 13, wherein the receiving the record to be appended is performed by one of the primary replica and the one or more secondary replicas that is closest to the sender.

15. The method of claim 14, further comprising:

forwarding the record to another one of the primary replica and the one or more secondary replicas that is closest to the one of the primary replica and the one or more secondary replicas that received the record from the sender.

16. A system for performing a record append operation, comprising:

means for receiving, by a primary server, a record append request that identifies a record to be appended to data stored by the primary server;

means for appending the record to the data;

means for forwarding the record append request to one or more secondary servers, the one or more secondary servers storing copies of the data; and

means for appending the record to the copies of the data by the one or more secondary servers.

17. A file system, comprising:

a master; and

a plurality of chunk servers connected to the master, one of the chunk servers, as a primary server, storing a chunk, at least one other one of the chunk servers, as at least one secondary server, storing a copy of the chunk,

the primary server being configured to:

receive a record append request that identifies a record to be appended to the chunk,

determine whether the record fits into the chunk,

append the record to the chunk when the record fits into the chunk, and

forward the record append request to the at least one secondary server,

the at least one secondary server being configured to:

write the record to the copy of the chunk.

18. The system of claim 17, wherein when determining whether the record fits into the chunk, the primary server is configured to determine whether appending the record to the chunk would cause the chunk to exceed a predetermined size.

19. The system of claim 17, wherein the primary server is further configured to pad the chunk to a predetermined size when the record does not fit into the chunk.

20. The system of claim 19, wherein the primary server is further configured to notify the at least one secondary server to pad the copy of the chunk to the predetermined size.

21. The system of claim 17, wherein the primary server is further configured to notify a sender of the record append request to retry the record append request on a next chunk when the record does not fit into the chunk.

22. The system of claim 17, wherein when forwarding the record append request, the primary server is configured to identify an offset at which the record was appended to the chunk.

23. The system of claim 22, wherein when appending the record to the copy of the chunk, the at least one secondary server is configured to write the record at the identified offset in the copy of the chunk.

24. The system of claim 17, wherein when appending the record, the primary server and the at least one secondary server are configured to write the record as an atomic unit.

25. The system of claim 17, wherein the primary server is configured to:

concurrently receive a plurality of record append requests, and

serialize the plurality of record append requests.

26. The system of claim 17, wherein the primary server is configured to concurrently process a plurality of record append requests.

27. The system of claim 17, wherein the master is configured to:

receive a request for identification of the primary server,

determine whether any of the chunk servers have a lease on the chunk, and

identify one of the chunk servers as the primary server when the one chunk server has a lease on the chunk.

28. The system of claim 27, wherein the master is further configured to:

grant a lease to one of the chunk servers when none of the chunk servers have a lease on the chunk.

29. The system of claim 17, wherein one of the primary sever and the at least one secondary server is configured to receive, from a sender of the record append request, the record to be appended, the one of the primary server and the at least one secondary server being closest to the sender.

30. The system of claim 29, wherein the one of the primary server and the at least one secondary server being configured to forward the record to another one of the primary server and the at least one secondary server that is closest to the one of the primary server and the at least one secondary server that received the record from the sender.

31. The system of claim 17, wherein when writing the record, the at least one secondary server is configured to overwrite existing data.

32. A method for performing a record append operation for a client in a system that includes a primary replica and one or more secondary replicas, comprising:

receiving, by the primary replica, a record append request from the client, the record append request identifying a record to be appended to a chunk stored by the primary replica, the one or more secondary replicas storing copies of the chunk;

determining whether if appending the record to the chunk would cause the chunk to exceed a predetermined size;

padding the chunk to the predetermined size when appending the record to the chunk would cause the chunk to exceed the predetermined size;

notifying the one or more secondary replicas to pad the copies of the chunk to the predetermined size; and

informing the client to retry the record append request on a next chunk.

33. A method for performing a record append operation, comprising:

receiving a record to be appended to a chunk;

receiving a record append request independent of the receiving of the record, the record append request identifying the record to be appended to the chunk; and

appending the record to the chunk as an atomic unit.

34. A file system, comprising:

a first server configured to:

store data,

receive a record to be appended to the data,

append the record to the data, and

forward the record to one or more second servers identifying an offset at which the record was appended to the data by the first server; and

each of the one or more second servers being configured to:

store copies of the data,

receive the record from the first server, and

write the record to the data at the identified offset.

35. The system of claim 34, wherein when writing the record, the one or more second servers are configured to append the record to the data.

36. The system of claim 34, wherein when writing the record, the one or more second servers are configured to overwrite a portion of the data with the record.

37. A method for performing record append operations for one or more clients in a system that includes a server, the method, performed by the server, comprising:

receiving a plurality of record append requests from the one or more clients, each of the record append requests identifying a record to be appended to data stored by the server;

serializing the record append requests to establish an order for the record append requests; and

appending the records to the data in the established order.

38. The method of claim 37, wherein the server is configured to concurrently perform one or more of the receiving, serializing, and appending for different ones of the record append requests.

39. A file system, comprising:

a first server configured to:

store data,

receive a record to be appended to the data, and

forward the record to a second one of the servers,

a third one of the servers being configured to:

store a copy of the data,

receive a record append request that identifies the record to be appended to the data,

append the record to the data as an atomic unit, and

forward the record append request to at least one of the first and second servers.